

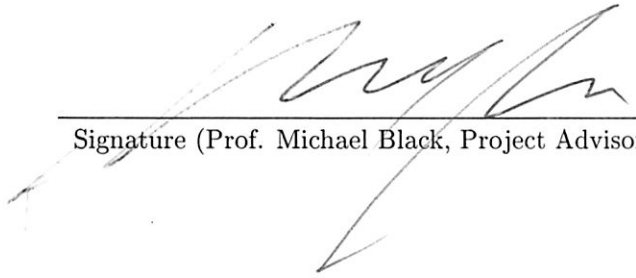
Decomposing Image Sequences into Layers According to Motion with the use of an Appearance Model

Robert Charles Altshuler

Department of Computer Science

Brown University

Submitted in partial fulfillment of the requirements for the Degree of Master of Science in the
Department of Computer Science at Brown University



Signature (Prof. Michael Black, Project Advisor)

5/14/03
Date

Abstract

Most techniques for motion analysis are based upon an assumption of spatial smoothness of optical flow. Strict enforcement of this constraint eases the task of computing the correct motion in some image regions, but also leads to erroneous estimations of motion in regions where the motion discontinuities are real. In fact, it is expected that such common occurrences as depth discontinuities, independently moving objects, and objects undergoing complex deformations will lead to violations of the smoothness constraint. As a result, techniques for motion analysis must employ special cases to predict the source of the motion discontinuity and compute the optical flow at the image locations where motion discontinuities occur.

We present a model for motion analysis that is based upon a robust estimation framework and makes use of a layered representation of images and flow fields to improve the estimation of multiple motions. The model associates different motions present in the image sequence with video layers and uses a system of weights to determine how those layers represent the motion at each pixel. Each layer also includes an image that models the appearance of that layer. The motion of a layer can be described either by an affine parameter set or a pre-computed basis flow field that allows the model to handle objects undergoing non-rigid deformations. In addition to motion estimation, the model is inherently able to perform segmentation of the image into layers. We present results from the analysis of two complex image sequences to demonstrate the model's capabilities.

1 Introduction

The fundamental principle on which most estimation of optical flow is based is that changes in image intensity in a small region during a short time interval are due only to the motion of objects in the scene relative to the camera. The constraint imposed by this principle, known as the brightness constancy assumption, makes the problem tractable, but also leads to some problems. Most notable is the aperture problem, that motion can only be computed in the direction perpendicular to the brightness gradient. Correspondingly, in regions of constant image intensity no motion can be computed.

Techniques for motion analysis commonly handle this problem by imposing another constraint on the spatial smoothness of optical flow. The spatial smoothness constraint is based upon an assumption that neighboring pixels represent the same object and therefore are undergoing similar motions. The spatial smoothness constraint suggests that the optical flow of image regions that lack texture can be interpolated from nearby areas where the brightness gradient is non-zero. The smoothness constraint also helps to reduce error from image noise in the motion estimated from the brightness constancy assumption.

Unfortunately, the smoothness constraint can not be enforced globally. The optical flow will not be smooth in such simple cases as one object moving in front of another, or even where surfaces at different depths move relative to the camera. The flow will be smooth across the surfaces of the objects, but at the object boundaries the flow will change rapidly. The problem, then, is that we must somehow detect the object boundaries and selectively not enforce the smoothness constraint

at them. We can go further and state that if we can accurately detect the object boundaries then we can take advantage of that information to improve our estimate of the optical flow.

We would like to be able to enforce spatial smoothness and improve our estimation of optical flow without having to explicitly detect object boundaries. We propose to accomplish this through the use of a layered model of optical flow. The layered model is implemented as a mixture model with the layers serving as the possible explanations of appearance changes in the observed images. Each layer accounts for one of the motions present in the image sequence. Smoothness is enforced by using an affine parameterization to describe the motion of a layer. A novel element of this layered model is the inclusion of an appearance model for each layer. This enables us to estimate motion between the images of the sequence and the layers and to analyze multiple images in parallel. The mixing probabilities of this mixture model are ownership weights for each pixel and each layer. The ownership weights associate pixels with layers according to how well the pixel's appearance is accounted for by the appearance model of a layer and the layer's motion.

We use an EM algorithm to optimize estimation of the motions of the layers and computation of the ownership weights. As part of the process of estimating the motion in the images, we also update the appearance of each layer. Updating the appearance models of the layers leads to a soft segmentation of the image according to the motions present in the image. This soft segmentation of the image in the appearance models of the layers of our model gives us the benefits of the knowledge of object boundaries without explicitly detecting them.

We recognize that motion discontinuities can also be caused by objects undergoing complex deformations. Consider, for example, the visual changes that can occur in an image sequence of a person's lips. As the lips move they can change shape dramatically, and significant appearance changes can occur as a result of the changes in the three dimensional shape of the lips. Aside from changes in appearance caused by changes in shape, in the case of skin, changes in appearance can also result from changes in color as the skin stretches or relaxes. In addition, for the case of a mouth, there is occlusion and disocclusion of the teeth and the mouth cavity as the mouth opens and closes.

While a person's lips are, perhaps, a particularly complex example, it is clear from this example that the optical flow from objects undergoing complex nonrigid deformations will lead to violations of the smoothness constraint. The motions of such objects can't be characterized using an affine parameterization. Accurate estimation of complex motions of this type requires an initial knowledge of the object and a model capable of using knowledge of the object's complex deformation to parameterize motion in the image sequence.

We handle the estimation of complex motions by introducing a variation of our layered model in which the motion of a layer is parameterized with a set of pre-computed basis flow fields. The basis flow fields are specific to the object in the scene and are computed using PCA so that they represent the most significant and varied components of the motions that the object undergoes. The observed motions of the object, and variations of these motions, can be produced as combinations of the basis flow fields. Again, the appearance models of the layers are an important element of the model. Each layer can have a very different appearance model, allowing different layers to account for very

different appearances of the object. We can then account for the complex motions of the object and the changes in appearance of the object by applying the motions produced as combinations of flow fields to the appearance models of the layers.

In the next section of this paper we review other layered models for image and motion analysis and discuss how they relate to our own model. The details of our model and its implementation are described in Section 3. In section 4 we present results from the analysis of two complex image sequences to demonstrate the model’s capabilities.

2 Related Work

A layered representation of images based upon the techniques of cel animation has been proposed by Adelson in [1]. In cel animation, images are painted on transparent sheets which are layered on top of each other and an opaque background to produce the animated scene. In Adelson’s layered representation, an image is composed of layers ordered by depth, just as in the case of cel animation. Each layer consists of an intensity map and an alpha map. The intensity map specifies the image for the layer. The alpha map specifies the transparency at each pixel and is used when compositing layers together. Adelson also proposes a velocity map for each layer which specifies how the pixels in the layer are moving and can be used to create an animated sequence from the layered representation.

We realize the utility of such a layered representation in motion analysis and include aspects of this representation in our model. Each layer in our model includes an image equivalent to the intensity map of Adelson’s representation. Layers also include motion parameters that provide information similar to that of the velocity map by specifying how the image of the layer can be warped to the observed image. We don’t composite the images as described in Adelson’s model, so rather than using an alpha map to indicate the contribution of a pixel to the composite image, we use ownership weights to represent the probability that the motion of the layer accounts for the appearance change of the pixel.

Wang and Adelson’s system for motion analysis [17, 18, 19] is based upon Adelson’s layered representation. Wang and Adelson’s system begins by computing the optical flow in an array of square image regions. These regions are pieced together using a k-means clustering technique and then each pixel is assigned to the motion estimate of exactly one of the cluster means. Regions where the motion estimate does not match the predicted motion are left unassigned. The process is repeated and the the motion estimates are iteratively refined. In the end, a set of motion estimates of minimum size are computed. Finally, the regions belonging to each motion estimate are assigned to layers and the layers are ordered by depth.

While Wang and Adelson’s system incorporates the concept of objects through the layered representation, it does not directly use the layer information to improve the motion estimation. Since the system is based upon an affine parameterization of motion, even if the optical flow of complex motions is correctly estimated in the initial array of square image regions, it is unlikely that these regions would be grouped together by the clustering algorithm. In the model that we present here,

we estimate the motion parameters of a layer as they apply to the entire image, not just small regions, and assign pixels to layers through ownership weights. Our model takes advantage of the layer information in motion analysis process by computing motion estimates between the observed images and the images of the layers. We also include a model parameterized by basis flow fields that can account for complex motions.

Recently, Frey and Jojic have published a series of papers [10, 11, 12, 15, 16] describing a probabilistic method for learning the appearance of objects in video layers of the type Wang and Adelson describe in their layered representation of images. Given a set of images and a set of transformations that can be applied in the compositing process, their method produces as output a set of layers that can be composited together to produce the observed images. Also, for each observed image they estimate the probability that each transformation is the correct transformation to apply to each layer so that when composited the observed image is produced.

The method presented by Frey and Jojic analyzes every choice of available transformations applied to each layer. This requires all of the transformations to be enumerated prior to analyzing the images and restricts analysis to only those transformations that have been enumerated. The enumeration of just every possible translation, rotation and scale quickly becomes intractable as the image size increases. To avoid this problem Frey and Jojic propose using the Fast Fourier Transform [16] as an alternate method for computing the likelihood of transformations being correct. However, implementing the FFT for this purpose limits translation to integer-pixel shifts. Also, using the FFT to examine rotations and scales requires transforming the image to a polar coordinate system, and as in the translational case the transformations would be discrete.

Similarly to Frey and Jojic, we compare the layers in our model to the images in the sequence instead of comparing the images in the sequence to each other. Also, like Frey and Jojic, we update the appearance of each layer as we analyze the image sequence. A significant difference between their model and ours is that we do not have an initial set of motions to choose from and must determine the motion in the image. However, by examining this problem from the standpoint of estimating optical flow, and by not compositing the layers, we are able to estimate continuous motion instead of being limited to estimating discrete motions, and we do not need to enumerate available transformations or switch coordinate systems.

Jepson and Black proposed the use of mixture models for motion analysis [14]. In the mixture model initially proposed, they allow for the multiple motions in an image patch examined by the model to be produced by a combination of motions occurring in that region. Multiple motions are treated as corresponding to layers, with each layer containing a single motion. Jepson and Black use an EM algorithm to iteratively compute probabilities that the motion constraint vectors at a particular image location belong to each layer and then update the layers. They also use robust regression techniques for motion estimation and include an outlier layer for constraint vectors that are not well accounted for by the motions of any other layers. The mixture model, however, is applied only to small image regions, not to entire images as is done in our model. It also makes no provision for estimating complex motions.

Black, Fleet and Yacoob have proposed another mixture model for motion analysis [8] similar to that of Jepson and Black. In this model, the changes in the image intensity are accounted for by a mixture of different models of appearance change. Among the models of appearance change included in this mixture are a model for changes due to object or camera motion, and another model for iconic changes. The model for iconic changes is intended to handle complex changes in appearance such as those due to occlusion, disocclusion or nonrigid deformations. A particular model for iconic change is specific to a certain object and must be precomputed by, for example, using an eigenspace representation of the object's appearance.

As in the model proposed by Jepson and Black, the different causes for appearance change are treated as layers, a robust regression technique is used and an EM algorithm is used to iteratively estimate the probability that a pixel belongs to each layer and then the parameters of the layers are updated. The model of iconic change used in the mixture model of Black, Fleet, and Yacoob can explain appearance changes due to complex motions, but because the model for iconic change is based upon eigenspace representations of appearance it does not produce a motion estimate corresponding to the appearance change. In the model we present here, we handle iconic change with a model based upon eigenspace representations of optical flow allowing us to account for the same types of complex appearance changes and also estimate motion.

3 A Layered Model of Optical Flow

The model for motion analysis that we present here is formulated as a mixture model. We construct the model as a set of layers that account for the appearance changes in the image sequence. Each layer includes an image that models the appearance of the layer and a set of parameters specifying the motion of the layer. In place of mixing probabilities, the model includes ownership weights associating pixels with layers that account for their appearance. So that we can make the best use of the information gained from the use of layers, we estimate motion between the image of each layer and an image in the sequence rather than estimating motion for consecutive images in the sequence. In applying the model to an image sequence we estimate the motion of each layer, compute the ownership weights for the pixels and also update the appearance model of each layer.

We construct the layered model out of two types of layers that handle different kinds of appearance changes; one for appearance changes due to affine motions and another for appearance changes due to more complex motions. In both cases, the model for a layer consists of an image of the layer, and weights indicating which pixels belong to that layer. Also, both models include parameters for the motion of that layer compared to each image in the sequence. In the affine model these parameters are used in equations that define the affine flow, while in the model for complex motions these parameters specify how the basis flow fields are combined.

The model also incorporates an additional outlier layer. Pixels that are not well accounted for by any particular layer of the model are assigned to the outlier layer. This allows us to avoid forcing a pixel which is not explained by the model for any layer into an incorrect layer where its inclusion

would lead to the erroneous estimation of parameters for that layer. As well, placing pixels into an outlier layer allows us to discover where our model is failing to account for the appearance change. By understanding where our model is failing we can improve the model, or develop additional models that can explain the appearance changes that the existing model can not.

As in [8], as we apply the model to a sequence of n images, we assume that an image $I(x, y, t)$ at location (x, y) and time t with $t = 1, \dots, n$ is generated by a combination of k layers, $I_l, l = 1, \dots, k$. A_l is the appearance model of layer l , and the image $I_l(x, y, t; \mathbf{a}_{l,t})$ is an image that has been generated by the motion of the observed image $I(x, y, t)$ as defined by the motion parameters, $\mathbf{a}_{l,t}$, of the layer l for that image.

According to the brightness constancy assumption, the changes in the appearance of an image at an image location (x, y) during a short time interval are due to the motion of the objects in the image. The brightness constancy assumption is described mathematically by the equation

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t). \quad (1)$$

We can expand the right hand side of this equation as a Taylor series and derive the equation

$$I_x u(\mathbf{a}) + I_y v(\mathbf{a}) + I_t = 0 \quad (2)$$

where I_x, I_y and I_t are the partial derivatives of I with respect to x, y and t , and $u(\mathbf{a})$ and $v(\mathbf{a})$ are functions of the motion parameters \mathbf{a} giving, respectively, the horizontal and vertical velocities of the pixel at location (x, y) . From this we can define the error of our model's estimates of u and v as the difference between the appearance model and the image produced by warping the observed image according to the motion parameters of the model.

$$\Delta I_l = I_x u(\mathbf{a}_{l,t}) + I_y v(\mathbf{a}_{l,t}) + I_t. \quad (3)$$

We then use this error function to compute the probability that the image was generated by the motion of the given layer. We use a robust likelihood function to compute this probability.

$$p_l(I(x, y, t) | \mathbf{a}_{l,t}, \sigma) = \frac{2\sigma^3}{\pi(\sigma^2 + \Delta I_l^2)^2} \quad (4)$$

Given the probability that the image was generated by each layer we can compute ownership weights to assign pixels to layers.

$$w_l(x, y, \sigma) = \frac{p_l(I(x, y, t) | \mathbf{a}_{l,t}, \sigma)}{\sum_{j=1}^k p_j(I(x, y, t) | \mathbf{a}_{j,t}, \sigma)} \quad (5)$$

This robust likelihood function falls off more rapidly than a gaussian, and causes appearance changes that are not well accounted for by the given layer to be accounted for by other layers or the

outlier layer. The σ parameter defines the breadth of the likelihood function. We can begin with a large value for σ allowing for a model to loosely fit many pixels, and then gradually decrease σ to narrow the likelihood function, forcing pixels to be more distinctly categorized into layers. Later, when we describe how the algorithm updates the ownership weights, we will discuss how pixels are assigned to the outlier layer.

We can then define the probability of an image being generated by the n layers as the mixture model given by

$$p(I(x, y, t) | \{\mathbf{a}_{l,t}, \sigma\}_{l=1}^k) = \sum_{l=1}^k w_l(x, y) p_l(I(x, y, t) | \mathbf{a}_{l,t}, \sigma). \quad (6)$$

If the ownership weights are shared between observed images, as is the case for the affine motion model, then the probability for the entire sequence of observed images can be written as

$$p(I(x, y, t) | \{\mathbf{a}_{l,t}, \sigma\}_{l=1}^k \{t\}_{t=1}^n) = \prod_{t=1}^n \sum_{l=1}^k w_l(x, y) p_l(I(x, y, t) | \mathbf{a}_{l,t}, \sigma). \quad (7)$$

This probability depends upon both the ownership weights and the motion parameters of the layers. In order to maximize the probability of an image being generated by the motion layers and not the outlier layer we must optimize both. We employ an EM algorithm to handle the simultaneous optimization of the ownership weights and the motion parameters. Updating the appearance models of the layers is done separately from estimation of the motion parameters and ownership weights. The update to the appearance model is computed as the mean of the warps of the images in the sequence according to the motion of the layer.

3.1 Affine Motion

In the affine model of motion presented here we can choose to use either the affine parameterization of optical flow or an extended planar representation. An affine transformation can be used to represent any combination of simple motions such as uniform scaling, translation, rotation, and shearing. As such, it is well known that an affine parameter set can be used to provide a good approximation of optical flow for surfaces undergoing rigid motion [3, 5, 13]. In an affine parameterization of optical flow, the equations for the horizontal flow, u , and vertical flow, v , are

$$\begin{aligned} u(x, y) &= a_1 + a_2x + a_3y \\ v(x, y) &= a_4 + a_5x + a_6y \end{aligned} \quad (8)$$

Similarly, in cases where the moving surface is slanted relative to the camera an 8-parameter planar representation of motion can be used [5, 13]. In the planar representation of optical flow the equations become

$$\begin{aligned}
u(x, y) &= a_1 + a_2x + a_3y + a_7x^2 + a_8xy \\
v(x, y) &= a_4 + a_5x + a_6y + a_7xy + a_8y^2
\end{aligned} \tag{9}$$

In order to estimate the parameters we perform a regression in which we warp the image in the sequence back to the image of the layer according to the current estimation of the parameters. Based on the derivatives between the warped image and the image of the layer we can compute an incremental update for the existing parameters. The motion estimate based on an affine or planar parameterization will vary smoothly across the image, so the spatial smoothness constraint is inherently enforced for each layer. The problem of determining whether to enforce the spatial smoothness constraint at motion discontinuities is handled through the use of the ownership weights. Pixels that are not well accounted for by the existing motion will have low ownership weights for the layer. When comparing the warped image and the image of the layer we take into account the ownership weights so that the update is not influenced by pixels that do not belong to the layer, such as those that are located at motion discontinuities.

3.2 Complex Motion Model

The complex motion model uses a precomputed set of basis flow fields to estimate the motion of objects undergoing complex nonrigid deformations. This, of course, requires a prior knowledge of the object in the image sequence. The basis flow fields can be produced by analyzing an image sequence of the known object with a more traditional method of motion analysis and then using PCA to produce a set of eigen flow fields representing the different motions that are present in the image sequence. If the image sequence to be analyzed with the complex motion model is of sufficient length then it can be used to produce the basis flow fields, if not, then an additional image sequence of the object must be used.

Given the horizontal and vertical flow fields, u and v , respectively, produced by the initial analysis, we create a column vector out of each flow field by ordering the flow values from left to right and top to bottom. We then construct two matrices, A_u and A_v , the columns of which are, respectively, the column vector representations of the u and v flow fields. A singular value decomposition of each of these matrices gives

$$\begin{aligned}
A_u &= U_u S_u V_u \\
A_v &= U_v S_v V_v
\end{aligned} \tag{10}$$

The matrices U_u and U_v are orthogonal and contain the eigenvectors of $A_u A_u^T$ and $A_v A_v^T$. Each of these eigenvectors is a flow field, and so we refer to them as eigen flow fields. The matrices S_u and S_v are diagonal and contain the singular values, which are ordered by magnitude. The magnitudes of the singular values also indicates the importance of their corresponding eigenvectors in representing the given flow fields.

We choose only a small number of the most important eigen flow fields to be the basis flow fields that we use to represent the motion fields with which we started. We can then produce reconstructions of the original flow fields, or other flow fields, by recombining these basis flow fields. For example, with $i = 1 \dots b$ horizontal basis flow fields, u_i , we can reconstruct a horizontal motion field as

$$u = \sum_{i=1}^b a_i u_i \quad (11)$$

where a_i is the reconstruction coefficient of the i th basis flow field. For a known motion field, the coefficient used to combine a basis flow field in the reconstruction is equal to the dot product of the basis flow field and the known motion field.

We’ve previously described some of the difficulties in analyzing complex motions, so it may seem odd to use a traditional method in producing our own model for complex motions. However, since we perform PCA on the motion field estimates of the traditional model the motion estimates from the traditional model do not need to be extremely accurate. The PCA algorithm will smooth out the errors and by definition, the eigen flow fields will be maximally different. So long as the eigen flow fields capture the various motions present in the sequence, the complex motion model should be able to produce accurate motion estimates by combining them correctly.

The complex motion model uses a linear combination of basis flow fields to produce estimates of motion in the image sequence. Each layer l is a combination of b basis flow fields. Each basis flow field specifies a value for the horizontal component of optical flow, u_b , and the vertical component, v_b , for every pixel in the image. The parameters for this model a are the reconstruction coefficients used to combine the basis flow fields. The parameter $a_{l,t,b}$ is the reconstruction coefficient for the b th basis flow field in layer l for the motion of the image observed at time t . A regression similar to that used by the affine model can be employed to incrementally update the parameters for the basis flow fields. The error between the warped image and the observed image is

$$\Delta I_{l,b} = \sum_b a_{l,b,t} (I_x u_b + I_y v_b) + I_t \quad (12)$$

where the update for each parameter $a_{l,b,t}$ for the layer is given by

$$\delta a_{l,b,t} = \frac{\sum_{x,y} w_l(x,y) (I_x u_b + I_y v_b) \Psi(\Delta I_{l,b}, \sigma)}{\sum_{x,y} w_l(x,y)^2 (I_x u_b + I_y v_b)^2} \quad (13)$$

This model also uses a separate set of ownership weights for each image in the sequence. This reflects the fact that the images for each layer may represent very different appearances of the object for which the motion is being estimated. For example, in the case of a mouth, the image of one layer may be a closed mouth while the image associated with another layer is an open mouth. In an image from the sequence displaying an open mouth, a pixel in a certain location may be part of

the mouth cavity and should have an ownership weight assigning it to the layer represented by the image of the open mouth. In another image from the sequence in which the mouth is closed, the pixel at the same location should be assigned to the layer with the image of the closed mouth.

3.3 Parameter Estimation

In applying our model to an image sequence we must estimate three parameters for the model; motion parameters, ownership weights and layer appearances. Estimates for the appearances of the layers can be initialized as selected images from the sequence to which the model is being applied. A greater effort must be made to estimate the motion parameters and solve for the ownership weights. The formulation of the layered model as a mixture model allows us to approach the problem of estimating the motion the parameters and solving for the ownership weights as one of maximum likelihood estimation.

We have previously given the formulation of the mixture model in equation 6. The probability of an image being represented by the layers of the model will be at a maxima when the partial derivative of the log likelihood, with respect to the motion parameters, is 0. So, as in [8], the maximum likelihood estimate of the motion parameters will satisfy the equation

$$\sum_{x,y} \sum_{l=1}^k w_l(x,y,\sigma) \frac{\partial}{\partial \mathbf{a}} \log p_l(I(x,y,t)|\mathbf{a}_{l,t},\sigma) = 0 \quad (14)$$

where the derivative of the log likelihood is given by

$$\frac{\partial \log p_l(I(x,y,t)|\mathbf{a}_{l,t},\sigma)}{\partial \mathbf{a}_{l,t}} = \Psi(\Delta I_l, \sigma) \frac{\partial I_l(x,y,t;\mathbf{a}_{l,t})}{\partial \mathbf{a}_{l,t}} \quad (15)$$

where

$$\Psi(\Delta I_l, \sigma) = \frac{-4\Delta I_l}{\sigma^2 + \Delta I_l^2} \quad (16)$$

We use an EM Algorithm to determine a maximum likelihood estimate of the motion parameters $\mathbf{a}_{l,t}$ and the ownership weights w_l for each layer. The EM algorithm is iterative, and in each iteration an Expectation Step and a Maximization step are performed. In the Expectation step, the weights are computed given the motion parameters as defined by equation 5. In the Maximization step the motion parameters are estimated based upon the current ownership weights. Rather than re-estimating the motion parameters with each iteration of the EM algorithm, we choose to improve incrementally upon the existing estimates. We can compute the incremental updates to the motion parameters using the regression technique appropriate for the layer's model of motion as described in sections (3.1) and (3.2).

As mentioned earlier, we can begin the execution of the EM algorithm with a relatively high value of σ , allowing very loose assignments of the pixels to layers. Gradually, as the motion parameters

improve, we can decrease the value of σ leading to stricter assignments of pixels to layers, and correspondingly, for each layer, estimates of the motion parameters that reflect only the pixels in the layer. The value for σ also determines the threshold for assigning a pixel to the outlier layer. If the sum of the probabilities for a pixel's appearance being explained with the motion parameterized layers of the model,

$$\sum_{l=1}^k p_l(I(x, y, t) | \mathbf{a}_{l,t}, \sigma)$$

is less than the value of the likelihood function at $\Delta I = 2.5\sigma$ then we assume the pixel is an outlier and set a weight of 1 for the outlier layer and for the other layers $w_l = 0$.

The appearance models of the layers are updated separately from estimation of the motion parameters and ownership weights. For the affine motion model, the update to the appearance model is computed as the mean of the warps of the images in the sequence according to the motion of the layer.

$$A_l = \frac{\sum_{t=1}^n I_l(x, y, t; \mathbf{a}_{l,t})}{n} \quad (17)$$

If a layer's appearance model and motion parameters properly account for the appearance of a pixel in the images of the sequence then the pixel will be stabilized. When the appearance model is updated, the stabilized pixels will appear in the updated appearance model as they appear in the observed images. Pixels that are not well accounted for will not be stabilized and will, instead, blur when the appearance model is updated. As the ownership weights associating pixels with layers are estimated, based upon the error in the warping of the observed images to the appearance model of the layer, the stabilized pixels will be accounted for by the appearance model of the layer and large ownership weights will be generated. However, blurred pixels in the appearance model will not account well for the pixels in the observed image leading to a small value for the ownership weight associating the pixel with the layer. This leads to the natural soft segmentation of the images in the sequence into the appearance models of the layers.

For the complex motion model, the update equation includes the ownership weights. We added ownership weights for each observed image to the complex model to account for the possibility of a pixel at a certain location to be explained by different appearance models at different times, as explained in section 3.2. For this same reason, when updating an appearance model, we don't want the update to include the pixel as it appears in observed images in which it is not explained by the appearance model being updated.

$$A_l(x, y) = \frac{\sum_{t=1}^n w_l(x, y, t) I_l(x, y, t; \mathbf{a}_{l,t})}{\sum_{t=1}^n w_l(x, y, t)} \quad (18)$$

In order to estimate large motions we estimate the model parameters at multiple image scales. We represent the multiple scales of the images with a Gaussian pyramid in which the resolution is

decreased by a factor of two at each level. We begin by estimating motion at the level with the lowest resolution. Once the motion parameters of the model have been estimated at a lower level we can exam the next higher resolution level using the motion parameters of the lower level as an initial estimate for the parameters of the higher level. The ownership weights and appearance models for the higher resolution level can both be initialized based on the initial motion parameters for that level.

At a given resolution level we can go through several iterations of first running the EM algorithm and then updating the appearance model. Repeating these steps several times allows for the appearance model to influence the estimation of motion parameters and ownership weights more so than if the appearance models are only updated during the initialization of parameters for the layer.

4 Results

4.1 The Pepsi Can Sequence

We begin by testing the layered model of optical flow on a sequence of seven images at a resolution of 201×201 pixels shown in Figure 1. The foreground of the scene in the images features a Pepsi can resting on a horizontal surface. The background of the scene is an artistic landscape of a mountain and the sky. The objects in the scene are stationary and the camera pans across the scene.

The panning of the camera produces several distinct motions in the image sequence. The soda can, in the foreground moves at one rate relative to the camera, while the background moves at a different, slower, rate. The apparent movement of the surface on which the can rests varies continuously as the distance of the surface from the camera changes. Of course, as the camera pans, the apparent motion of the soda can leads an occlusion boundary between the can and the table, and another between the can and the background. Corresponding disocclusion boundaries are also present. Additionally, the surface of the can is reflective. As is commonly the case with reflective objects, the reflections and specularities on the surface of the can do not move along with the can. Instead, the reflections and specularities shift across the surface of the can changing shape and brightness.

The image sequence was analyzed with two layers using an affine motion model, with the intention that one layer would account for the background and another for the can. The variation in depth over the surface of the table is significant compared to the distance of the table from the camera, so a third layer was not used as it was not expected that the apparent motion would be accurately described even if a planar parameter set was used. Figure 2 shows a larger version of the middle image of the sequence. The middle image of the sequence was used to initialize both of these layers and motion was analyzed between each layer and each of the remaining six images.

We know that the motion in this image sequence is caused only by a panning motions of the camera, so we enforced a restriction on the affine parameters to limit the motion to translations. Also, since we expect to be able to correctly estimate the motion of most of the pixels we imposed

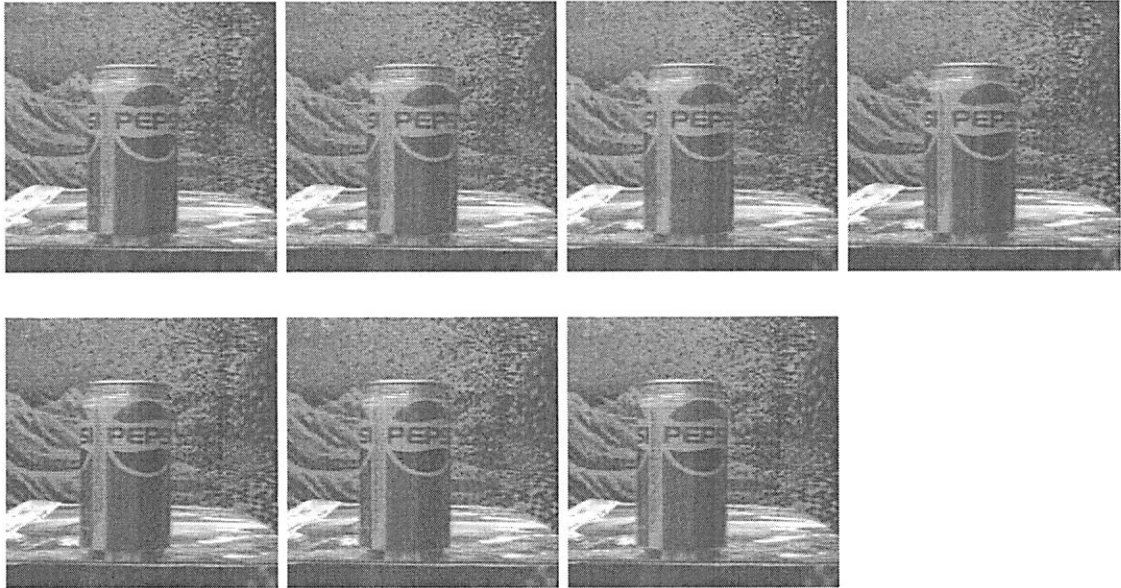


Figure 1: The Pepsi can image sequence

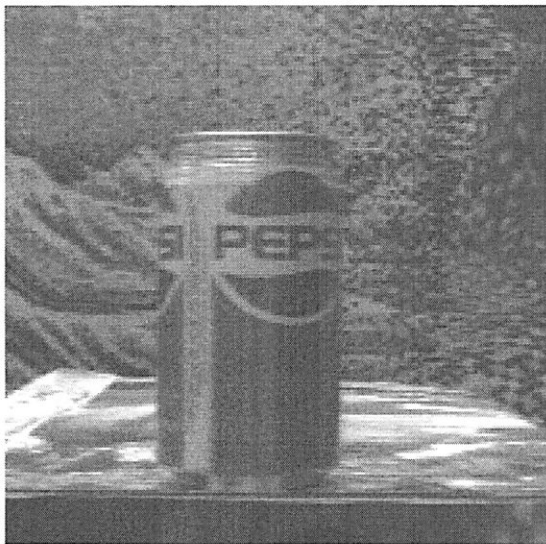


Figure 2: The middle image of the sequence is used to initialize the layers

a restriction for the ownership weights of outliers to be split between the other layers rather than assigned to the outlier layer. The images were analyzed at 3 resolution levels and with 30 iterations of the EM algorithm at each scale. The appearance model was updated once at each scale.

Figure 3 shows images representing the ownership weights for the layers. These images clearly show the segmentation of the observed images. The ownership weights for layer 1 are brightest, and therefore have the highest values, for the background. The ownership weights for layer 2 are complementary to that and are highest for the pixels of the Pepsi can and lowest for the background. The ownership weights for layer 2 are also very high for some of the pixels of the table at approximately the same depth as the base of the can. There are some areas on the surface of the can where the ownership weights indicate that the pixels are associated evenly with both layers. This corresponds to the regions on the cans surface that lack texture. Since the motion of the background is less than the motion of the can, the motion of these areas can be equally well accounted for by both layers.

Although the ownership weights are not directly involved in the estimation of the appearance model, because the ownership weights indicate a segmentation of the image into layers there should also be a segmentation in the appearance models of the two layers, shown in Figure 4, for the reasons explained in section 3.3. As we would expect, in the appearance model for layer 1 the mountains in the background are sharp while the Pepsi can is blurred, particularly in the area of the letters and where there are vertical edges. The appearance model for layer 2 also agrees with the ownership weights, containing a sharp and unblurred image of the Pepsi can, while the background is blurred, most noticeably in the area to the left of the Pepsi can.

Finally, we can visually confirm that the estimated motions are correct by stabilizing the observed images. The stabilized images can be generated by warping the observed images according to the motion parameters for a layer. The motion parameters characterize the motion of the observed image compared to the appearance model of the layer, so at locations where the ownership weights for the layer are high the appearance of the warped image should match the appearance model of the layer. If the motion parameters are estimated correctly then the warped images should match each other in areas accounted for by the same layer. Stabilized images for the first and last images in the sequence are shown in figure 5. Comparison of the stabilized images shows that in the images stabilized with the motion parameters of layer 1 the background remains still while the can moves and in the images stabilized using the motion parameters of layer 2 the can is stable and the background moves. This agrees with our expectations for the given layer appearance models and ownership weights and confirms that the model has accurately estimated the motion for the images.

4.2 A Human Mouth

We tested the complex motion model on a human mouth. A set of eight horizontal and eight vertical basis flow fields, shown in figure 6 were chosen to define the set of possible motions for the model. The basis flow fields were computed by analyzing a sequence of nearly 3500 images of a mouth opening, closing, and displaying a number of facial expressions.

The model was tested on a sequence of 7 images in which the mouth is initially partly open and

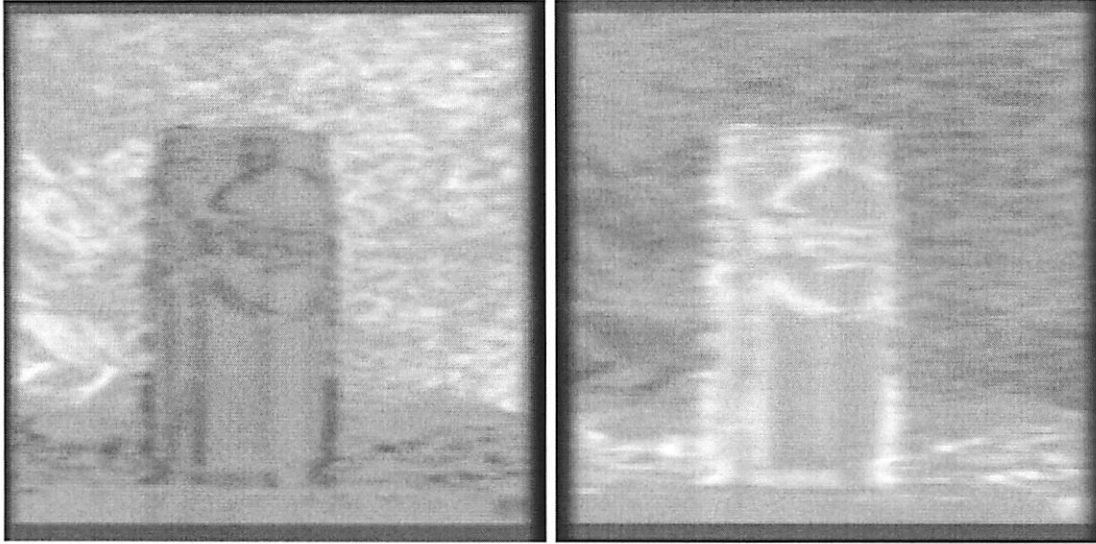


Figure 3: The ownership weights for layer 1 (left) and layer 2 (right)



Figure 4: The appearance models for layer 1 (left) and layer 2 (right)

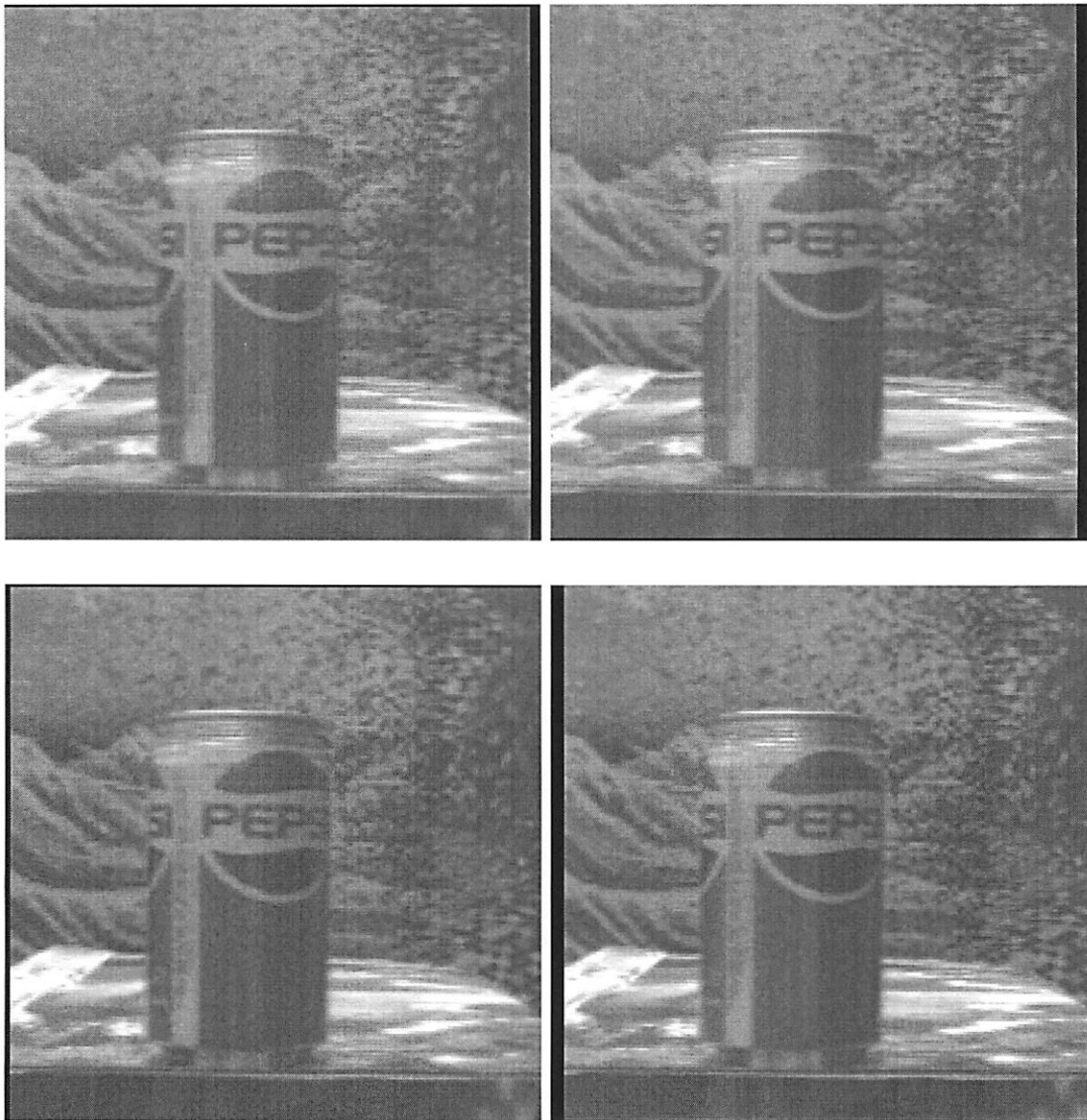


Figure 5: The stabilized images for layer 1 (left) and layer 2 (right)

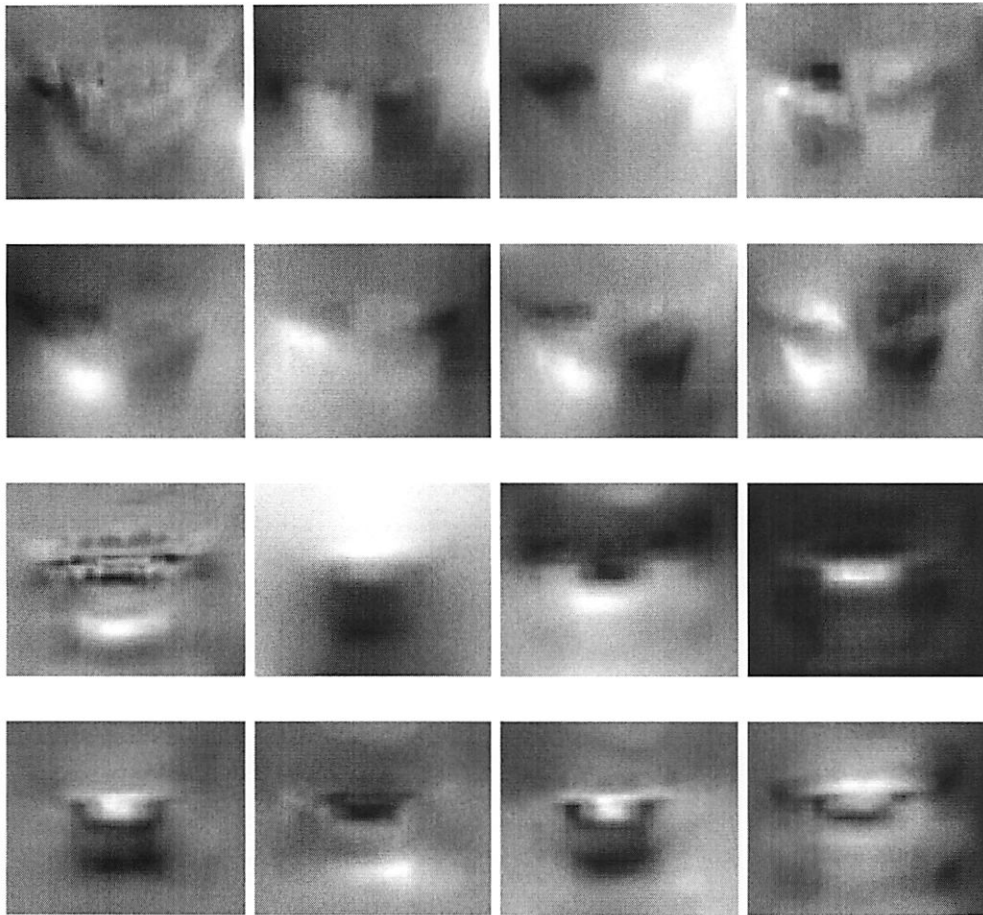


Figure 6: Horizontal (top two rows) and vertical (bottom two rows) basis flow fields for a complex motion model for a mouth.

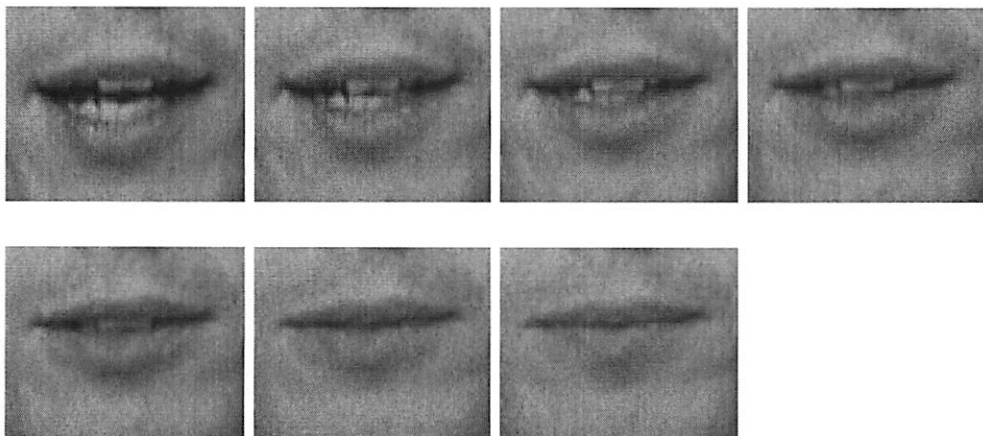


Figure 7: The images of the mouth sequence analyzed using the complex motion model

gradually closes (fig. 7). Over the course of the image sequence the shape of the lips changes and the mouth cavity and teeth disappear. Aside from the non-rigid motion of the lips the disappearance of the mouth cavity and teeth is an appearance change that many other motion analysis techniques are not well suited to handle.

The layered model was constructed with two layers to account for the opened and closed mouth. The first image of the sequence was used to initialize the first layer and the last image of the sequence was used to initialize the second layer. The remaining 5 images were analyzed using the model. The model was tested with a pyramid with 3 resolution levels and one execution of the EM algorithm and update to the appearance models per level.

The ownership weights for the layers are shown in figures 8 and 9. The ownership weights are fairly evenly split between the layers for many of the pixels that are outside the area of the lips and the top lip. This makes sense since these the images used to initialize the appearance models appear nearly the same for these regions. For the images at the beginning of the sequence in which the mouth is open the ownership weights are high for the first layer in the areas of the lower lip, teeth and mouth cavity. The ownership weights for the first layer gradually decrease and for the latter images of the sequence the ownership weights are low over most of these areas. The opposite trend is present in the ownership weights for layer 2. Meanwhile, the weight images for the outlier layer (fig. 10) show that in several of the images a small number of pixels, mostly along the crease of the lips, are not being accounted for well by either layer.

The updated appearance models are shown in figure 11. The updated appearances both seem to have changed toward a partly open mouth, which makes sense given that over the entirety of the analyzed image sequence the mean appearance of the mouth is partly open. The appearance model for the first layer unquestionably retains the appearance of a mouth, although it is slightly less open than in the initial appearance model, with the teeth closed and the mouth cavity no longer visible.

The updated appearance model for the second layer is more interesting. The bottom lip has shifted downwards and is vertically stretched and a faint image of teeth have appeared. It may seem that the updated appearance of the second layer is simply a mix of the observed images in which the mouth is partly open. If that were the case we would expect the area near the corners of the lips would be darker in the updated mean, as it is in the most of the images of the sequence, and it is not. This demonstrates the influence of the ownership weights in computing the appearance update for this layer. The addition of the teeth in this appearance model may explain why some pixels along the crease of the lips have been assigned to the outlier layer.

It is possible to generate stabilized images for the mouth sequence just as was done when testing the affine motion model. The stabilized images for the mouth sequence are shown in figure 12. For the first layer the bottom lip has been stabilized for the first four images of the sequence. As the mouth closes the lower lip is stretched downwards to match the appearance model of the layer. The stabilized images for the second layer are more subtle, but the bottom lip appears stabilized for the last three images of the sequence and the top lip seems to be stabilized for the last four images.

The stabilized images do demonstrate a shortcoming of this model. The model estimates motion

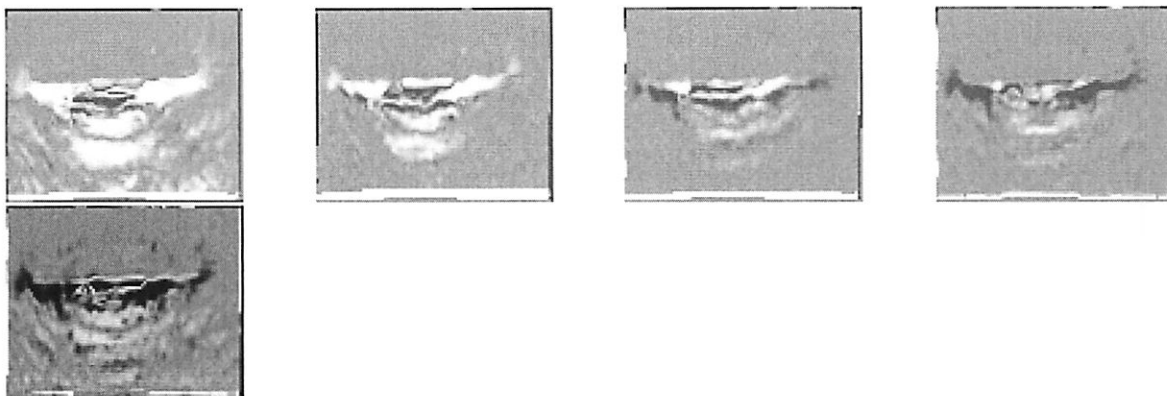


Figure 8: The ownership weights for each image of the mouth sequence for layer 1

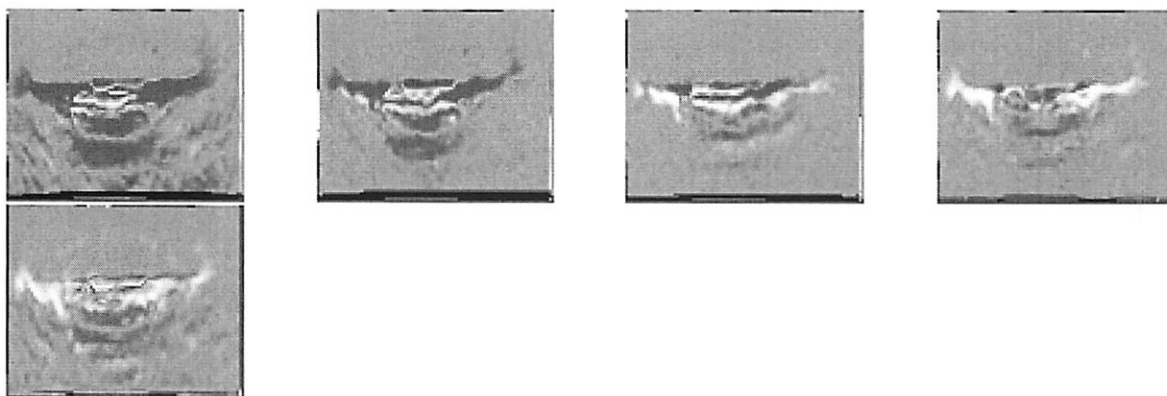


Figure 9: The ownership weights for each image of the mouth sequence for layer 2

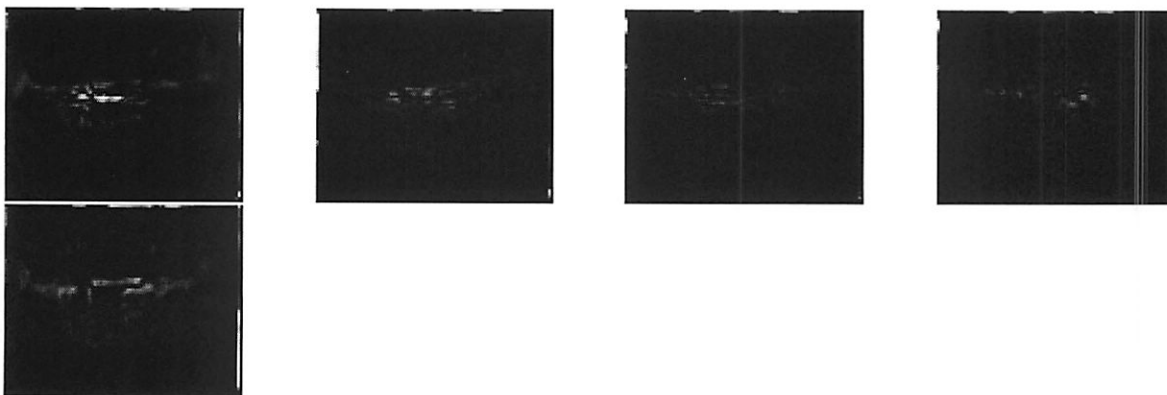


Figure 10: The ownership weights for each image of the mouth sequence for the outlier layer

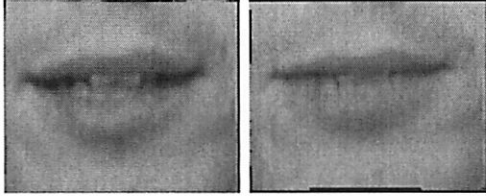


Figure 11: The updated appearance models of the layers in the complex motion model

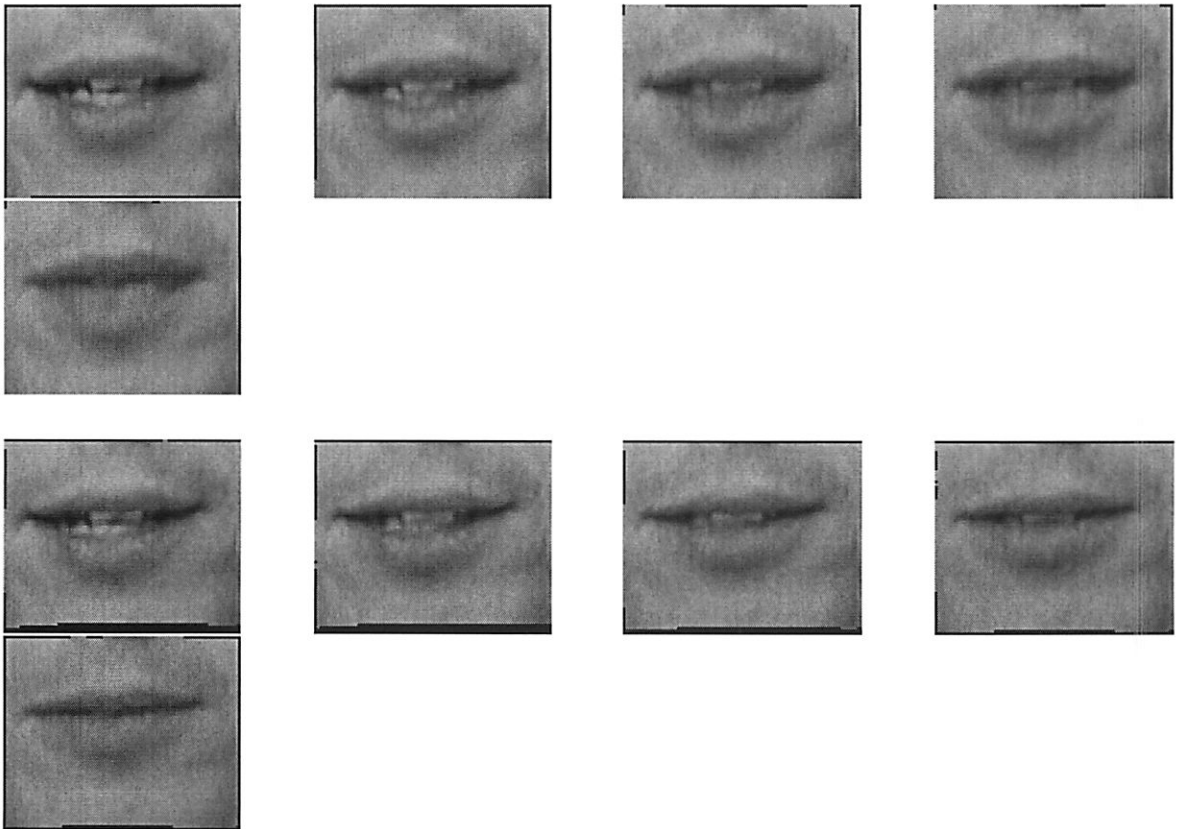


Figure 12: The images of the mouth sequence stabilized according to the motion parameters for layer 1 (top) and layer 2 (bottom)

by performing a backwards warping; it warps the observed image towards the layer’s appearance model. This simplifies the formulation of the model, but the warped images can not match match the appearance model if the appearance model contains pixels that are occluded in the observed image. More generally, a warp can handle occlusion of pixels, but not disocclusion of pixels. In this example, the appearance model includes teeth, but when warping an observed image of a mostly closed mouth, in which the teeth are occluded, toward that of an open mouth it is not possible for the teeth to be recreated in the warped image. The use of a forward warping, in which the appearance model is warped toward the observed image, could resolve this problem and is an opportunity for future improvement.

We can also examine the vertical flow fields (fig. 13) to confirm that the motion estimates are correct. The flow fields for the second layer all indicate a downward movement of the pixels in the top half of the image and an upward movement of the pixels in the lower half of the images with larger movements for the first three images of the sequence. This makes perfect sense since those motions would cause an image of an open mouth to close. The exact opposite trend is present for all of the images for the first layer, except for the first image, although that may simply indicate that the mouth appears slightly more closed in the appearance model than in the first image.

5 Conclusion

We have developed a model for motion analysis that takes advantage of a layered representation of images and optical flow to improve the estimation of motion. The model makes use of an appearance model for each layer allowing for multiple images to be analyzed at once. The observed images can then be compared to the appearance model of each layer, and the appearance model of each layer can be updated so that each represents only the portions of the image whose appearance change is accounted for by the motion of the layer. The model is also capable of using a set of basis flow fields specific to an object in the image sequence to handle the estimation of complex motions.

A number of opportunities exist for future improvement to the model presented here. A technique for detecting object location into the complex motion model to allow for the basis flow fields to be translated and centered on the object whose motion they represent. The layered model could also be modified to warp the appearance model of the layer toward the image rather than warping the image toward the appearance model as is done in the current model. A system for determining the depth ordering of the layers could allow for new images to be generated from the appearance models of the layers of the model. Future work could also include methods for allowing the appearance model to change over the course of long image sequences.

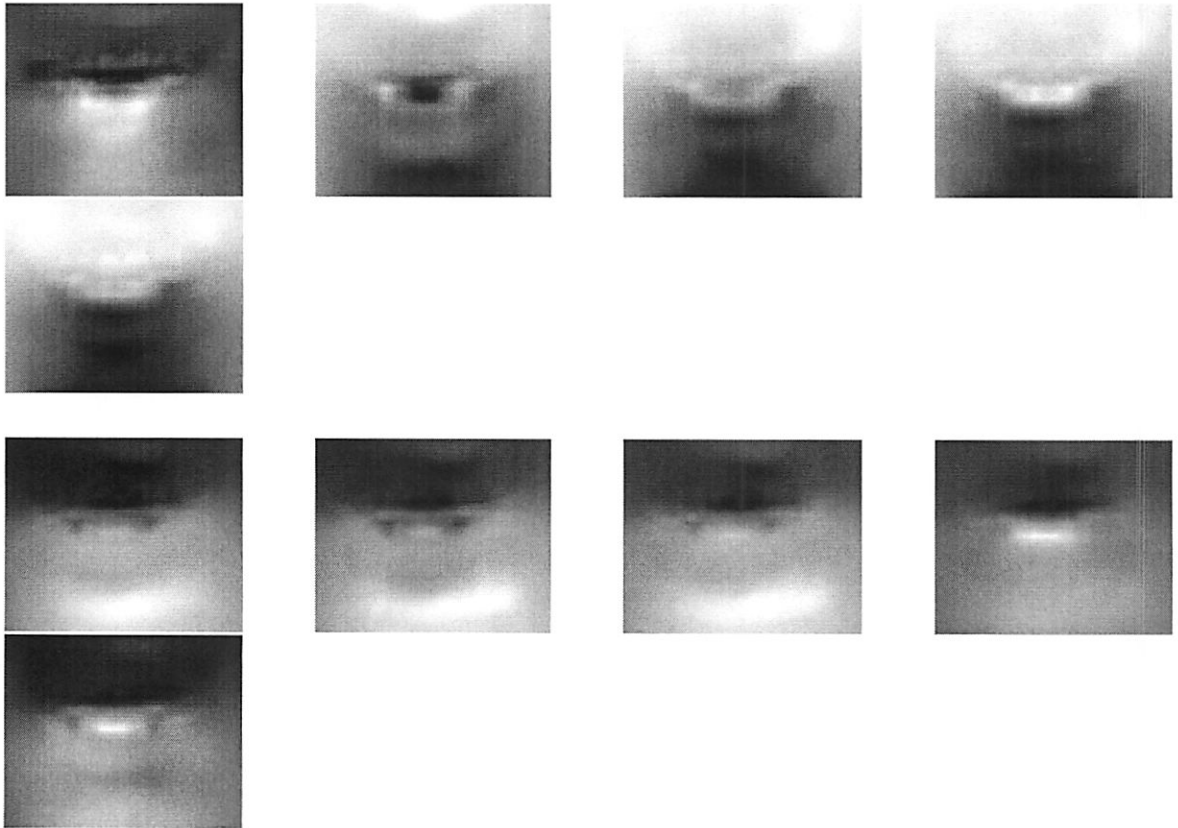


Figure 13: The vertical flow fields estimated for the images of the mouth sequence for layer 1 (top) and layer 2 (bottom)

References

- [1] E.H. Adelson. *Layered Representation for Image Coding*. Technical Report No. 181, Vision and Modeling Group, MIT Media Lab, December 1991.
- [2] E.H. Adelson and P. Anandan. *Ordinal Characteristics of Transparency*. AAAI Workshop on Qualitative Vision, pp.77-81. July 20, 1990, Boston, MA
- [3] G. Adiv. *Determining Three-Dimensional Motion and Structure from Optical Flow Generated by Several Moving Objects*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 7(4):384-401, July 1985
- [4] P. Anandan. *A Unified Perspective on Computational Techniques for the Measurement of Visual Motion*. In International Conference on Computer Vision, pages 219-230. London, May 1987
- [5] J.R. Bergen, P. Anandan, K.J. Hanna. R. Hingorani. *Hierarchical Model-Based Motion Estimation*. In European Conference on Computer Vision, 1992, pp. 237-252
- [6] M.J. Black. *Robust Incremental Optical Flow*. PhD. Thesis, Yale University, New Haven, CT, 1992. Research Report: YALEU/DCS/RR-923
- [7] M.J. Black and P. Anandan. *Robust Estimation of Multiple Motions: Parametric and Piecewise-Smooth Flow Fields*. Computer Vision and Image Understanding 63, p.75-104, 1996
- [8] M.J. Black, D.J. Fleet and Y. Yacoob. *Robustly Estimating Changes in Image Appearance*. Computer Vision and Image Understanding 78, p.8-31, 2000
- [9] P.J. Burt and E.H. Adelson. *The Laplacian Pyramid as a Compact Image Code*. IEEE Transactions on Communication, 31:532-540, 1983
- [10] B.J. Frey and N. Jojic. *Estimating Mixture Models of Images and Inferring Spatial Transformations Using the EM Algorithm*. IEEE Conference on Computer Vision and Pattern Recognition, p. 416-422, June 1999
- [11] B.J. Frey and N. Jojic. *Transformation-Invariant Clustering and Dimensionality Reduction Using EM*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Nov. 2000
- [12] B.J. Frey and N. Jojic. *Transformation-Invariant Clustering Using the EM Algorithm*. Accepted to IEEE Transactions on Pattern Matching and Machine Intelligence.
- [13] B.K.P. Horn. *Robot Vision*. The MIT Press, Cambridge, MA, 1986
- [14] A. Jepson and M.J. Black. *Mixture Models for Optical Flow Computation, in Partitioning Data Sets: With Applications to Psychology, Vision and Target Tracking (Ingmer Cox, Pierre Hansen, and Bela Julesz, Eds.)*. DIMACS Workshop, April 1993, pp. 271-286, Amer. Math. Soc., Providence, RI.

- [15] N. Jojic and B.J. Frey. *Learning Flexible Sprites in Video Layers*. IEEE Conference on Computer Vision and Pattern Recognition, 2001
- [16] N. Jojic and B.J. Frey. *A Generative Model for 2.5D Vision: Estimating Appearance, Transformation, Illumination, Transparency and Occlusion*. Submitted to International Journal on Computer Vision, 2002
- [17] J.Y.A. Wang and E.H. Adelson. *Layered Representation for Motion Analysis*. Proceedings of the IEEE Computer Vision and Pattern Recognition Conference. pp.361-366, New York, June 1993
- [18] J.Y.A. Wang and E.H. Adelson. *Spatio-Temporal Segmentation of Video Data*. Proceedings of the SPIE: Image and Video Processing II, vol. 2182, San Jose, February 1994
- [19] J.Y.A. Wang and E.H. Adelson. *Representing Moving Images with Layers*. IEEE Transactions on Image Processing Special Issue: Image Sequence Compression, vol 3, no. 5, p.625-638, September 1994